Multiple-Site Ligand Binding to Flexible Macromolecules: Separation of Global and Local Conformational Change and an Iterative Mobile Clustering Approach

VELIN Z. SPASSOV,[1,2] DONALD BASHFORD[1]

[1] *Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, California 92037*
[2] *Institute of Biophysics, Bulgarian Academy of Sciences, Sofia, Bulgaria*

Received 5 November 1998; accepted 4 March 1999

ABSTRACT: This article concerns the calculation of equilibria of ligand binding to multiple sites in macromolecules in the presence of conformational flexibility and conformation-dependent interaction among the sites. A formulation of this problem is presented in which global conformational changes are distinguished from conformational changes that are confined to ''locally flexible regions.'' The formalism is quite general in that ligands of different types, multivalent binding sites, tautomeric binding sites, and sites that bind more than one type of ligand can be accommodated. Strictly speaking, the separation of the conformational problem into global and local parts does not impose any loss of generality, although in practice it is necessary to restrict the number of global and local conformers. Because of the combinatorics of binding and conformational states, the computational complexity of a problem having only local conformational flexibility grows exponentially with the number of sites and the number of locally flexible regions. An iterative mobile clustering method for cutting off this exponential growth and obtaining approximate solutions with low computational cost is presented and tested. In this method, a binding site is selected, and a ''cluster'' of strongly interacting sites is set up around it; within the cluster, the binding and conformational states are fully enumerated, whereas the influences of sites outside the cluster on the sites inside are treated by a mean field approximation. The procedure then moves to the next site around which another

Correspondence to: D. Bashford; e-mail: bashford@scripps.edu
Contract/grant sponsor: National Institutes of Health
Contract/grant number: GM45607

CCC 0192-8651 / 99 / 111091-21

(possibly overlapping) cluster is formed and the calculation is repeated. The procedure iterates through the list of sites in this way, using the results of previous iterations for the mean-field terms of current iterations until a convergence criterion is met. The method is tested on a large set of randomly generated problems of varying size, whose geometries are chosen to have protein-like statistical properties. It is found that the method is accurate and rapid with the computational cost scaling linearly to quadratically with the number of sites, except for a minority of cases in which large clusters occur by chance. The new method is more accurate than a Monte Carlo method, and may be faster or slower depending on the clustering criteria and details of the macromolecule. © 1999 John Wiley & Sons, Inc. J Comput Chem 20: 1091–1111, 1999

Keywords: multiple site; titration; pK_a; ligand building; mean field approximation; conformational flexibility; electrostatics

Introduction

Biological macromolecules typically have multiple sites at which protons or other ligands can bind or dissociate. These molecules also have a considerable degree of conformational flexibility, which may be energetically coupled to ligand binding. The characteristic example is the pH titration of proteins, during which several tens, or even hundreds of side chains are able to take up or release protons, and may simultaneously undergo conformational changes through side-chain motions, local loop motions, or even global conformational changes. It may be desired to calculate average properties, such as the protonation fraction of a particular site, as functions of some measure of an external chemical potential, such as pH, or to follow the populations of various conformers as functions of the external chemical potentials. This can be done, in principle, by calculating statistical ensembles over all possible binding and conformational states of the macromolecule. If it can be assumed that the energetics of ligand binding can be decomposed into terms intrinsic to the sites and site–site interaction terms, then the overall problem can be divided into two stages: (1) determination of these individual terms; and (2) the combinatorics involved in the enumeration of all possible states of the molecule. The scope of this study includes the combinatorial problem.

The combinatorial problem can be divided into three levels: ligand binding; localized conformational change; and global conformational change. Ligand binding to multiple sites leads to a number of binding states that increases exponentially with the number of sites. For example, if N_s sites can each be in one of two binding states, there are 2^{N_s} overall binding configurations possible. Localized conformational change leads to similar exponential growth. The number of conformers grows like the number of conformers per local region, raised to the power of the number of local regions. Finally, we can include local conformational changes such as hinge-bending or allosteric transitions, which may be energetically coupled to ligand-binding changes at many sites throughout the molecule. Although these global conformational changes may not combine with one another in a way that leads to exponential growth, they do introduce energetic couplings between binding events at distant sites in a way that might defeat many strategies for dealing with multisite combinatorics through the neglect of interactions or correlations at long distance.

In this study we first present a general formalism that includes all three levels of the multisite, multiconformational problem. It has no restrictions as to the number of different kinds of ligands binding to the macromolecule, so its is applicable to problems such as the interplay of protonation and allosteric transition in hemoglobin, or the effects of pH on the cooperativity of Ca^{2+} binding by calbindin.[1] It has no restrictions on the number of binding states that an individual site may have, nor does it require that each site bind only one type of ligand. Therefore phenomena such as tautomerism, polyvalent acids, and coupled redox-protonation events fit naturally into the formalism.

We then present an approximation method based on designating a cluster of strongly interacting sites, enumerating all states within this subset of sites explicitly, and treating the effects of the

sites outside the cluster by a mean-field method. This approximation amounts to a neglect of correlations between weakly coupled sites. The calculation proceeds by iteration to self-consistency. On each iteration, each of the binding sites is taken in turn and used as the center of a cluster. The mean-field terms for sites outside the cluster are taken from averages calculated on the previous iteration. The combinatorics of binding and of local conformational change are handled simultaneously. We call this method iterative mobile clustering (IMC). Global conformational change cannot be dealt with directly within this method, but if the number of global conformers is not too large, separate IMC runs can be made for each one, and they can be combined using a generalization of a standard formula involving integrals of ligand-binding occupancies over chemical potential. This accounts properly for any extra correlations between sites introduced by couplings to global conformational changes.

Finally, we present calculations to test the accuracy and efficiency of the method. Because we wish to make statistical tests of the method's performance over large numbers of macromolecular systems, and because the realistic determination of energetic terms, such as intrinsic pK values and site–site interactions, is beyond the scope of this study, we have tested the method on highly simplified systems for which the generation of various ''molecules'' and the calculation of energy terms is very rapid. However, we have chosen the functional forms and parameters of these test systems to provide combinatorial problems of a similar level of difficulty to those expected in real proteins.

BACKGROUND

In the earliest approaches to this problem, the drastic simplifications of the physical model, and interest in the total titration curve, rather than individual-site titration curves, made it possible to use analytical approximations to solve the multisite problem. For example, Linderstrøm-Lang[2] used a model in which all titrating sites of a given kind were assumed to have identical intrinsic properties and to be smeared over the surface of a sphere so that all site–site interactions had the same value, W. It is then possible to show that, in the limit of a large number of titrating groups, the midpoint of the total titration curve is offset by an amount proportional to $W\overline{Z}$, where $\overline{Z}$ is the average charge due to the titrating groups. Tanford and Kirkwood[3] also assumed identical intrinsic properties, but interactions were derived from a model of point charges arranged symmetrically within a sphere. The partition functions of the proton configuration can then be expanded to obtain approximate expressions for an average effective interaction. More detailed models became possible as structures became available and computer power increased, but most calculations did not include explicit conformational change. Tanford and Roxby[4] used a model in which each site had its own unique set of interactions with other sites, and introduced a set of equations based on the intuitively reasonable approximation that the effective pK of each site at a particular pH was shifted by the site's interaction with the *average* charge of the other sites. These mean-field type equations can be derived formally by assuming an uncorrelated form for the distribution function.[5] Bashford and Karplus[5] showed that this neglect of correlations can lead to significant errors in the presence of strong site–site interactions and introduced an alternative method in which the sites that could not alter their protonation significantly at a particular pH were regarded as fixed, and the exact partition sum was calculated for the reduced set of variable sites. The method is highly accurate but is usually limited to molecules with fewer than 30 or 40 sites because the reduced number of sites is roughly half the original number. Beroza et al.[6] overcame this limitation by developing a Monte Carlo method whose computational cost scales approximately as the square of the number of sites. Gilson [7] developed a clustered mean-field approach in which the partition sums of groups of strongly interacting residues are calculated in the presence of the mean field of groups not in the cluster. Yang et al.[8] used a similar method except that clusters were defined around each site as the calculation proceeded instead of being fixed in advance. We refer to this variation as ''mobile clustering.''

You and Bashford[9] introduced explicit conformational variability into protein titration calculations, but the conformational change was limited to side-chain movements and their effects were limited to the intrinsic titration properties of the moving site. The interplay of conformational change and the multisite problem was neglected to avoid the combinatorial problem. Several calculations have been presented in which the multisite, multiconformer problem is made tractable by imposing rather severe limitations on the conformational changes considered. Beroza and Case[10] used

an extension of the previous Monte Carlo method to study a system in which each titrating side chain was allowed to have two conformations, and simplifying assumptions as to their interaction were introduced. Alexov and Gunner[11] limited conformational change to the positions of titrating and/or hydrogen bonding protons. Formalisms in terms of global conformational change have been presented and used in calculations based on molecular dynamics.[7,12–15]

Theory

GENERAL FORMULATION OF PROBLEM

The applicability of the methods developed in this study depends mainly on the energetic couplings between the ligand-binding sites of the macromolecule, and not on the type of ligands bound or the intrinsic chemical properties of the binding sites. Therefore, the formalism will be developed at first in a very general way.

Consider a molecule, M, that has N_s sites at which ligands may bind. The ligands may be protons, electrons, Ca^{2+} ions, etc. Different sites on the same molecule may bind ligands of different types, and the same site may even bind different ligands (e.g., the ubiquinones of photosynthetic reaction center undergo both protonation/deprotonation and oxidation/reduction). The overall binding state of M may be denoted by an N_s-component vector $\mathbf{x}$ whose elements x_i specify the binding state of the ith binding site. The values that a particular x_i may take on depend on the particular characteristics of the ith site; for example, a lysine side chain might only have the x_i values ''protonated'' and ''deprotonated,'' whereas a histidine side chain might have the values ''δ-tautomer,'' ''ϵ-tautomer,'' and ''protonated.'' A ubiquinone's possible x_i values might be ''protonated/oxidized,'' ''protonated/reduced,'' ''deprotonated/oxidized,'' and ''deprotonated/reduced.'' Denoting the ligand types as $L_1, L_2, \ldots,$ and their chemical potentials as $\mu_1, \mu_2, \ldots,$ consider the reaction:

$$M(\mathbf{x}) \leftrightarrow M(\mathbf{0}) + \nu_1(\mathbf{x})L_1 + \nu_2(\mathbf{x})L_2 + \cdots$$

where $\mathbf{0}$ is a reference binding state of M and $\nu_1(\mathbf{x})$ is the number of molecules of species L_1 bound in forming $M(\mathbf{x})$ from the reference state. The equilibrium of this reaction relates the relative activity of M($\mathbf{x}$) to the chemical potentials of the ligands and the standard chemical potentials of M in different binding states:

$$\mu_M(\mathbf{x}) = \mu_M(\mathbf{0}) + \mu_1\nu_1(\mathbf{x}) + \mu_2\nu_2(\mathbf{x})\ldots$$

$$-RT\ln\frac{[M(\mathbf{x})]}{[M(\mathbf{0})]} = \mu^\circ_M(\mathbf{x}) - \mu^\circ_M(\mathbf{0}) - \mu_1\nu_1(\mathbf{x}) - \mu_2\nu_2(\mathbf{x})\ldots$$

where the chemical potential formulas for the dilute macromolecule, $\mu_M = \mu^\circ_M + RT\ln[M]$, have been used in the second step. This can be generalized to an expression for the fractional population or probability of finding the macromolecule in binding state $\mathbf{x}$:

$$f(\mathbf{x}) = Z^{-1}e^{-\beta[\mu^\circ_M(\mathbf{x}) - \mu_1\nu_1(\mathbf{x}) - \mu_2\nu_2(\mathbf{x})\ldots]}$$

where $\beta = (RT)^{-1}$, and Z is a normalization constant assuring that the sum of $f(\mathbf{x})$ over all possible binding states, $\mathbf{x}$, is equal to 1. For ordinary ligands, such as protons, the ligand activity enters though the usual definitions such as, $\mu_{H+} = \mu^\circ_{H+} - 2.3RT$ pH. (We have written ''2.3'' for the value of $\ln 10$ throughout to save clutter in the formulas.) Redox processes can be incorporated by formally regarding electrons as ''ligands'' having a chemical potential $\mu_e = -FE$, where F is the Faraday constant and E is the electrochemical potential.[16]

The macromolecule may also undergo transitions between a finite number, N_c, of distinct conformers. These transitions may be thought of as unimolecular ''reactions'' and incorporated into the chemical potential so that the probability of finding the molecule in conformer c with binding state $\mathbf{x}$ is:

$$f(c,\mathbf{x}) = Z^{-1}e^{-\beta[u^\circ_M(c,\mathbf{x}) - \mu_1\nu_1(\mathbf{x}) - \mu_2\nu_2(\mathbf{x})\ldots]} \quad (1)$$

where Z now includes a summation over the conformers.

BINDING SITE INTERACTION AND LOCAL CONFORMATION

In this subsection, we will build up the formulation for the problem by first introducing the functional forms related to ligand binding, then those related to conformational change, and finally the combination of the two, and the resulting expression for the probability of a particular conformational/binding state. We also introduce the distinction between local and global conformational change through the concept of local flexible regions (LFRs).

Assume that the dependence of the chemical potential as a function of binding states is a sum of one- and two-site terms:

$$\mu_M^\circ(\mathbf{x}) = \text{const.} + \sum_i^{N_s} P_i(x_i) + \frac{1}{2}\sum_{ij}^{N_s}{}' W_{ij}(x_i, x_j)$$

where the notation Σ' indicates that the diagonal terms, $i = j$, are excluded from the summation. The P are related to the intrinsic binding affinity of the individual sites (the relation to ''intrinsic pK_a'' for the proton-binding case is discussed in what follows), and the W are the binding-state-dependent site–site interactions (e.g., electrostatic interaction between ionizable groups). There may be small groups of three or more closely associated sites whose binding energetics cannot be represented in this way; for example, a redox-active metal center with two pH-titratable ligands. These groups may be regarded as a single extended site, because our formalism does not restrict the number of states or the types of ligands that can be represented by the binding vector elements, x_i.

We will consider two types of conformational change: global, such as allosteric transitions or unfolding of proteins; and localized, such as side-chain rotations or loop rearrangements. No special restrictions are placed on the former, but the latter are assumed to be confined to N_r *local flexible regions* (LFRs). The LFRs are defined in such a way that the dependence of the chemical potential is a sum of one- and two-LFR terms, as detailed in what follows. Thus, the demarcation of LFRs is closely related to assumptions about the form of the chemical potential function. If the integer index k labels the N_g global conformers, and $\mathbf{c}$ is a vector of N_r elements, c_a, which label the conformers of the ath LFR, the conformational energy is then:

$$E(k, \mathbf{c}) = E_g(k) + \sum_a^{N_r} E_a(k, c_a) + \frac{1}{2}\sum_{ab}^{N_r}{}' V_{ab}b(k, c_a, c_b)$$

The energetics of conformation and binding are assumed to be coupled as follows: A binding site ''belongs'' to one and only one LFR, but an LFR may contain multiple binding sites. The one-site binding energies, P_i, are influenced by the global conformation and the conformation of the LFR to which it belongs, but not by the conformation of any other LFR. The site–site interactions, W_{ij}, are influenced by the global conformation and the conformations of the LFRs to which the binding sites i or j belong, but not by any other LFRs. The expression for chemical potential as a function of both binding state and conformational state is then:

$$\begin{aligned}\mu_M^\circ(k, \mathbf{c}, \mathbf{x}) = E_g(k) + \sum_a^{N_r}\Bigg[& E_a(k, c_a) + \sum_{i\in s_a}^{N_s} P_i(k, c_a, x_i) \\ & + \frac{1}{2}\sum_{i,\, j\in s_a}{}' W_{ij}(k, c_a, x_i, x_j)\Bigg] \\ & + \frac{1}{2}\sum_{ab}^{N_r}{}'\Bigg[V_{ab}(k, c_a, c_b) \\ & + \sum_{\substack{i\in s_a \\ j\in s_b}} W_{ij}(k, c_a, c_b, x_i, x_j)\Bigg]\end{aligned} \tag{2}$$

where s_a denotes the subset of binding sites belonging to LFR a.

Perhaps the simplest example of an LFR and binding state in a protein is a flexible, pH-titratable side chain. The LFR might consist of the side chain, and its states, c_a, might be a set of specifications of its dihedral angles. This LFR would contain only one binding site whose state, x_i, is either ''protonated'' or ''deprotonated.'' The P_i energy term for the site would pertain to its intrinsic titration properties, and its W_{ij} terms to its interactions (probably electrostatic) with other binding sites. The designation of this side chain as an LFR and site would imply that changes in its binding state and conformation have no effect on the interaction of other pairs of sites or LFRs. A slightly more complex example might be a flexible backbone loop bearing two pH-titratable side chains as well as other nontitrating residues. In this case, it would probably be necessary to consider the whole loop as a single LFR containing two sites, because it would be impossible to express the loop's flexibility in terms of independent variations associated with the two titratable residues. A still more complex case might be the Ca^{2+} binding loop of an ''EF-hand'' protein such as calbindin. This could include, for example, a flexible loop, a Ca^{2+} binding site and two aspartic acid residues. At first this might appear to be an LFR containing three binding sites, but it may be that the strong interactions between the carboxylates and the Ca^{2+} ligand would not be expressible in the pairwise additive form of eq. (2). It would then be necessary to merge the three sites into a single ''extended site''

capable of binding two protons and one Ca^{2+} ion in various combinations.

Before substituting this into the expression for the probability of a given state, we note that the ligand terms of that expression, such as $\mu_1\nu_1(\mathbf{x})$, can always be expressed as a sum of one-site terms, such as $\sum_i \mu_1\nu_{1i}(x_i)$, where $\nu_{1i}(x_i)$ is the number of ligands of type 1 bound to site i (relative to a reference state) when it is in binding state x_i. Our full expression for the probability of finding the macromolecule in the conformation indicated by the global index k and the local indices $\mathbf{c}$ and in the ligand-binding state $\mathbf{x}$ is then:

$$\begin{aligned} f(k,\mathbf{c},\mathbf{x}) &= Z^{-1}\exp\Bigg\{-\beta\Bigg[E_g(k)+\sum_a^{N_r}\Bigg[E_a(k,c_a) \\ &\quad+\sum_{i\in s_a}P_i(k,c_a,x_i)-\mu_1\nu_{1i}(x_i)\cdots \\ &\quad+\frac{1}{2}\sum_{i,j\in s_a}{}'\,W_{ij}(k,c_a,x_i,x_j)\Bigg] \\ &\quad+\frac{1}{2}\sum_{ab}^{N_r}{}'\Bigg[V_{ab}(k,c_a,c_b) \\ &\quad+\sum_{\substack{i\in s_a\\ j\in s_b}}W_{ij}(k,c_a,c_b,x_i,x_j)\Bigg]\Bigg]\Bigg\} \end{aligned} \tag{3}$$

where:

$$\begin{aligned} Z &= \sum_k^{N_g}\sum_{\mathbf{c}}\sum_{\mathbf{x}}\exp\Bigg\{-\beta\Bigg[E_g(k)+\sum_a^{N_r}\Bigg[E_a(k,c_a) \\ &\quad+\sum_{i\in s_a}^{N_s}P_i(k,c_a,x_i)-\mu_1\nu_{1i}(x_i)\cdots \\ &\quad+\frac{1}{2}\sum_{i,j\in s_a}{}'\,W_{ij}(k,c_a,x_i,x_j)\Bigg] \\ &\quad+\frac{1}{2}\sum_{ab}^{N_r}{}'\Bigg[V_{ab}(k,c_a,c_b) \\ &\quad+\sum_{\substack{i\in s_a\\ j\in s_b}}W_{ij}(k,c_a,c_b,x_i,x_j)\Bigg]\Bigg]\Bigg\} \end{aligned} \tag{4}$$

It will be useful to separate the problems of global and local changes by introducing $f(\mathbf{c},\mathbf{x}\mid k)$, the conditional probability of finding the molecule in the local conformation and binding state denoted by $\mathbf{c}$ and $\mathbf{x}$, provided that the global conformational state is k. It is defined in terms of the full probability distribution through the formulas:

$$f_g(k)=\sum_{\mathbf{c}}\sum_{\mathbf{x}}f(k,\mathbf{c},\mathbf{x}) \tag{5}$$

$$f(\mathbf{c},\mathbf{x}\mid k)=\frac{f(k,\mathbf{c},\mathbf{x})}{f_g(k)} \tag{6}$$

The definition implies that the conditional probability function is normalized to one.

COMPLEXITY OF PROBLEM

The main difficulty in using this formula is the summation over states in the calculation of the partition function, Z. As an example, consider a protein with N_s pH-titrating sites that can be in either a protonated or deprotonated state. Suppose that each titrating site forms its own LFR that can be in three different conformational states (e.g., corresponding to side-chain rotamers), and that there are N_g global conformations. The number of states is then $N_g \times 3^{N_s} \times 2^{N_s}$. The exponential growth in the number of states with the number of sites is a serious problem because typical proteins have several tens, or even hundreds, of titrating sites.

Another potential problem is the determination of the functions, such as P_i and W_{ij}, which make up the chemical potential. In the aforementioned example, there are N_s functions, P_i, whose arguments, k, c_a, and x_i can take on $N_g \times 2 \times 3$ different values. Therefore, a complete tabulation of all possible values of all of the P functions requires $6N_sN_g$ actual P-function evaluations. A tabulation of all of the site–site coupling functions, W, would require approximately $N_g(6N_s)^2$ evaluations. Similar reasoning applied to the other terms shows that the growth of problem of tabulating the chemical potential terms is, at worst, quadratic in the number of sites, quadratic in the number of binding or conformational states, and linear in the number of global conformers. This problem is not trivial (see Discussion), but it is not as severe as the exponential growth in the number of states.

The main focus of this study is on approximation methods that cut off the exponential growth of the number of states in the partition sum with the number of sites and LFRs in the macro-

molecule. Such methods will be useful only if N_q is small enough that explicit summation over the N_g global conformers is manageable. In other words, the conformational variability of the molecule may be very great, but it must come about mostly through the combinatorics of relatively small sets of local conformational changes.

MEAN FIELD APPROACHES

In mean-field approaches to statistical mechanical problems, the influence upon a subsystem of the rest of the system is treated in an average way so that enumeration of the states of the rest of the system can be avoided. Such approximations can be formally derived by assuming a form of the distribution function f that contains no correlations between the subsystem and the rest of the system, writing the free energy (or other appropriate thermodynamic potential) as a functional of this uncorrelated distribution f, and minimizing the functional with respect to f under the constraints of normalization.[17] Here we introduce the approximation that a small cluster of sites or LFRs can be regarded as uncorrelated with the rest of the system, and derive a mean-field approximation using the free energy functional approach. (This should not be confused with the ''cluster field approach'' of Dimitrov and Crichton.[18]) The calculation for the cluster grows exponentially with the number of sites in the clusters, but does not grow rapidly with the total number of sites in the molecule, so that, if the cluster can be made sufficiently small, the calculation is tractable. Gilson has described a similar strategy for the case of pH titration without conformational change.[7] We then describe the iterative mobile clustering method in which clustering is applied in turn at each site.[8] These methods are formulated in a very general way, making no new assumptions about the nature of the ligand binding or the relation of LFRs to binding sites, apart from the neglect of certain correlations. We then specialize to the case of pH titration with flexible side chains, which is the basis of the test model calculations. Finally, we discuss a limiting case of the cluster method, which is analogous to the Tanford–Roxby[4] method, and the relationship between clustering, mobile clustering, and the global conformational problem.

Clustering Approximation

Before taking the functional variation approach to deriving a cluster-based mean-field approximation, we will confirm our choice of the thermodynamic potential functional by showing that the exact expression (1) can be obtained from the variational method when the distribution is allowed to remain fully general. Formally, the distribution (1) corresponds to a semigrand canonical ensemble in which each system of the ensemble contains one macromolecule whose energy levels are given by $\mu^\circ_M(c, \mathbf{x})$ while the number of ligand molecules varies. The corresponding thermodynamic potential is $\langle \mu^\circ_M \rangle - \langle \nu_1 \rangle \mu_1 - \langle \nu_2 \rangle \mu_2 \cdots - TS$. This can be written as a functional of the distribution f:

$$F[f(c, \mathbf{x})] = \sum_{c, \mathbf{x}} f(\mu^\circ_M(c, \mathbf{x}) - \nu_1 \mu_1 \cdots + \beta^{-1} \ln f) \tag{7}$$

where the second term is based on the Gibbs entropy postulate,[19] $S = -R\langle \ln f \rangle$, and conformers are labeled by c, as in eq. (1), which includes both local and global flexibility. Introducing the Lagrange multiplier, $A + \beta^{-1}$, to enforce the normalization of f, and setting the functional derivative to zero gives:

$$\frac{\delta}{\delta f} \sum f(\mu^\circ_M - \nu_1 \mu_1 \cdots + \beta^{-1} \ln f) + (A + \beta^{-1})(1 - f) = \mu^\circ_M - \nu_1 \mu_1 \cdots + \beta^{-1} \ln f - A = 0,$$

leading to:

$$f = e^{\beta A} e^{-\beta[\mu^\circ_M - \nu_1 \mu_1 \cdots]}$$

which is the distribution expected. The normalization condition gives:

$$e^{-\beta A} = \sum_{c, \mathbf{x}} e^{-\beta[\mu^\circ_M - \nu_1 \mu_1 \cdots]}$$

from which we see that A, which was introduced in the Lagrange multiplier, is the corresponding equilibrium free energy.

Let us select a *cluster* of LFRs that interact strongly with one another but whose correlations with other LFRs are to be neglected either because the interactions are weak or, in the case of strong interactions involving some ''peripheral'' LFRs of the cluster, because we are interested primarily in the more ''central'' LFRs. We denote this cluster of LFRs as Γ. Sites belonging to the LFRs within Γ are also within the cluster; in other words, the boundary of the cluster cannot separate binding sites that are within the same LFR. We now introduce the essential approximation: *that the condi-*

tional probability function (6) *contains no correlations across the boundary of the cluster*. The functional form of the probability function is therefore:

$$f_g(k) f_\Gamma(\mathbf{c}, \mathbf{x} \mid k) f_{\text{out}}(\mathbf{c}', \mathbf{x}' \mid k) \tag{8}$$

where the vectors appearing in f_Γ are reduced vectors containing only elements pertaining to sites and LFRs within the cluster Γ, and the primed vectors appearing in f_{out} pertain only to sites and LFRs outside of the cluster Γ. Note that, because the product form above neglects correlations only in the *conditional* probability, overall correlation between sites inside and outside the cluster can still arise through couplings to the global conformational changes denoted by k. This is desirable. It means, for example, that the titration of a protein residue that triggers global unfolding can be correlated to titration events in distant regions of the protein.

The free energy functional is obtained by substituting eq. (8) for the distribution function in eq. (7). Using the normalization properties of f_Γ and f_{out}, the energy terms can be grouped according to whether they involve sites within the cluster, sites outside the cluster, or interactions between sites inside and outside the cluster. To simplify, only one ligand species is shown, but the methods are also applicable to multiple ligand species. Including the Lagrange multiplier terms for the f_g, f_Γ, and f_{out}, the free energy functional is:

$$\begin{aligned}
&F[f_g, f_\Gamma, f_{\text{out}}] \\
&= \sum_k^{N_g} f_g(k) \Bigg\{ E_g(k) + \sum_{\mathbf{c}} \sum_{\mathbf{x}} f_\Gamma(\mathbf{c}, \mathbf{x} \mid k) \\
&\quad \times \Bigg[\sum_{a \in \Gamma} \Bigg(E_a(k, c_a) + \sum_{i \in s_a} P_i(k, c_a, x_i) \\
&\quad - \mu_1 \nu_{1i}(x_i) + \frac{1}{2} \sum_{i,\, j \in s_a}{}' W_{ij}(k, c_a, x_i, x_j) \Bigg) \\
&\quad + \frac{1}{2} \sum_{a,\, b \in \Gamma}{}' \Bigg(V_{ab}(k, c_a, c_b) \\
&\quad + \sum_{\substack{i \in s_a \\ j \in s_b}}{}' W_{ij}(k, c_a, c_b, x_i, x_j) \Bigg) \Bigg] \\
&\quad + \sum_{\mathbf{c}'} \sum_{\mathbf{x}'} f_{\text{out}}(\mathbf{c}', \mathbf{x}' \mid k) [\text{similar terms outside } \Gamma] \\
&\quad + \sum_{\mathbf{c}} \sum_{\mathbf{x}} \sum_{\mathbf{c}'} \sum_{\mathbf{x}'} f_\Gamma(\mathbf{c}, \mathbf{x} \mid k) f_{\text{out}}(\mathbf{c}', \mathbf{x}' \mid k) \\
&\quad \times \Bigg[\sum_{\substack{a \in \Gamma \\ b \notin \Gamma}} \Bigg(V_{ab}(k, c_a, c_b') \\
&\quad + \sum_{\substack{i \in s_a \\ j \in s_b}} W_{ij}(k, c_a, c_b', x_i, x_j') \Bigg) \Bigg] \\
&\quad + \beta^{-1} \sum_{\mathbf{cx}} f_\Gamma \ln f_\Gamma \\
&\quad + \beta^{-1} \Bigg[\sum_{\mathbf{c}'\mathbf{x}'} f_{\text{out}} \ln f_{\text{out}} + \beta^{-1} f_g \ln f_g \Bigg] \Bigg\} \\
&\quad + \Big(A_g + \beta^{-1} \Big) \Big(1 - \sum_k f_g(k) \Big) \\
&\quad + \sum_k \big(A_\Gamma(k) + \beta^{-1} \big) \Big(1 - \sum_{\mathbf{cx}} f_\Gamma(\mathbf{c}, \mathbf{x} \mid k) \Big) \\
&\quad + \sum_k \big(A_{\text{out}}(k) + \beta^{-1} \big) \Big(1 - \sum_{\mathbf{c}'\mathbf{x}'} f_{\text{out}}(\mathbf{c}', \mathbf{x}' \mid k) \Big)
\end{aligned} \tag{9}$$

Setting the functional derivative with respect to f_Γ to zero leads to the following expression for f_Γ:

$$\begin{aligned}
f_\Gamma(\mathbf{c}, \mathbf{x} \mid k) = e^{\beta A_\Gamma(k)} \exp \Bigg\{ -\beta \Bigg[&\sum_{a \in \Gamma} \Bigg(E_a(k, c_a) \\
&+ \sum_{i \in s_a} P_i(k, c_a, x_i) - \mu_1 \nu_{1i}(x_i) \\
&+ \frac{1}{2} \sum_{i,\, j \in s_a}{}' W_{ij}(k, c_a, x_i, x_j) \Bigg) \\
&+ \frac{1}{2} \sum_{a,\, b \in \Gamma}{}' \Bigg(V_{ab}(k, c_a, c_b) \\
&+ \sum_{\substack{i \in s_a \\ j \in s_b}} W_{ij}(k, c_a, c_b, x_i, x_j) \Bigg) \\
&+ \sum_{\substack{a \in \Gamma \\ b \notin \Gamma}} \Bigg(\langle V_{ab}(k, c_a) \rangle_b \\
&+ \sum_{i \in s_a} \sum_{j \notin s_b} \big\langle W_{ij}(k, c_a, x_k) \big\rangle_{bj} \Bigg) \Bigg] \Bigg\}
\end{aligned} \tag{10}$$

where the influence of sites and LFRs outside Γ appears through the mean-field terms:

$$\langle V_{ab}(k, c_a)\rangle_b = \sum_{c'_b} f_b(c'_b \mid k) V_{ab}(k, c_a, c'_b) \quad (11)$$

$$\left\langle W_{ij}(k, c_a, x_i)\right\rangle_{bj} \equiv \sum_{c'_b}\sum_{x'_j} f_{bj}(c'_b, x'_j \mid k) W_{ij}(k, c_a, c'_b, x_i, x'_j) \quad (12)$$

where:

$$f_{bj}(c'_b, x'_j \mid k) \equiv \sum_{\mathbf{c}'\text{ex.}c'_b}\ \sum_{\mathbf{x}'\text{ex.}x'_b} f_{\text{out}}(\mathbf{c}', \mathbf{x}' \mid k) \quad (13)$$

$$f_b(c'_b \mid k) \equiv \sum_{x_j} f_{bj}(c'_b, x'_j \mid k) \quad (14)$$

Normalization through the Lagrange multiplier $A_\Gamma(k)$ requires:

$$e^{-\beta A_\Gamma(k)} = \sum_{\mathbf{c}\in\Gamma}\sum_{\mathbf{x}\in\Gamma} \exp\left\{-\beta\left[\sum_{a\in\Gamma}\left(E_a(k, c_a) + \sum_{i\in s_a} P_i(k, c_a, x_i) - \mu_1\nu_{1i}(x_i) + \frac{1}{2}\sum_{i,j\in s_a}{}' W_{ij}(k, c_a, x_i, x_j)\right) + \frac{1}{2}\sum_{a,b\in\Gamma}{}'\left(V_{ab}(k, c_a, c_b) + \sum_{\substack{i\in s_a\\ j\in s_b}} W_{ij}(k, c_a, c_b, x_i, x_j)\right) + \sum_{\substack{a\in\Gamma\\ b\notin\Gamma}}\left(\langle V_{ab}(k, c_a)\rangle_b + \sum_{i\in s_a}\sum_{j\in s_b}\left\langle W_{ij}(k, c_a, x_i)\right\rangle_{bj}\right)\right]\right\} \quad (15)$$

It can be seen that $A_\Gamma(k)$ has the form of a free energy of the cluster, Γ, in the mean field of the rest of the macromolecule in global conformer, k. If Γ has been chosen so that the number of LFRs and sites within it is not too large, the evaluation of the above expressions for f_Γ will be practical, but the mean field terms, or the one-site and one-LFR distributions for the sites outside the cluster, must be known. An iterative procedure for obtaining these will be described in the next subsection.

An expression for f_{out} entirely analogous to eqs. (10)–(15) can be obtained by differentiation of the free energy functional [eq. (9)] with respect to f_{out}. In general, the number of sites outside the cluster is not small enough to make the direct use of such formulas tractable.

An expression for f_g can be obtained by differentiation of the free energy functional with respect to f_g. The immediate result is a complex expression in terms of the f_Γ and f_{out}, but it can be simplified by using the Lagrange multipliers and the normalization conditions on f_Γ and f_{out}. One finally obtains:

$$f_g(k) = e^{\beta A_g}\exp\left\{-\beta\left[E_g(k) + A_\Gamma(k) + A_{\text{out}}(k) - \sum_{\substack{a\in\Gamma\\ b\notin\Gamma}}\left(\langle V_{ab}(k)\rangle_{ab} + \sum_{\substack{i\in s_a\\ j\in s_b}}\left\langle W_{ij}(k)\right\rangle_{aibj}\right)\right]\right\} \quad (16)$$

This result can be understood intuitively as the canonical distribution corresponding to a Hamiltonian made of the global energy function, E_g, and the free energies of the subsystems inside and outside of Γ.[7] The mean-field terms entering with a negative sign are necessary to compensate for double counting, because similar mean-field terms are present in both A_Γ and A_{out}. The direct use of eq. (16) will not be practical in general, because the number of sites outside the cluster is not small enough to allow for the calculation of A_{out}. However, in later subsections we examine the case of single-site clusters in which similar expressions can be evaluated; in addition, we present an approach to the global conformational ensemble that can be used when the direct approach is not possible.

Iterative Mobile Clustering (IMC)

Generally, one is interested not in finding the full distribution function $f(k, \mathbf{c}, \mathbf{x})$, but in calculating average quantities that only require much simpler distributions. For example, calculating the degree of protonation of a site or of the entire macromolecule requires only the one-site distribu-

tions, eqs. (13) and (14). If we limit our goal to the calculation of these distributions we can avoid the problem of calculating f_{out} by the following iterative scheme.

Each iteration involves a pass through the list of the macromolecule's sites (or LFRs). Taking each site (or LFR) in turn as the ''central'' site, establish a cluster Γ around that site by the criteria that a site will be included in Γ if the magnitude of its energy of interaction with the central site (as given by the W function) is greater than a cutoff energy. Use eqs. (10) and (15) to find the f_Γ for this cluster, using the results of previous iterations for the one-site or one-LFR averages. Having obtained f_Γ, calculate the one-site and one-LFR distribution functions for the central site, and save them for use in the next iteration. The iteration stops when a self-consistency requirement has been met.

One drawback of mobile clustering is that there is no longer a single unique free energy functional of the distribution, such as eq. (9). This makes it impossible to write an expression such as eq. (16) for f_g. However, we will see that the f_g problem can be approached in another way.

pH Titration with Flexible Side Chains

The test systems to be used here are meant to mimic proteins with flexible pH-titrating side chains interacting electrostatically. The attribution of local flexibility to side chains implies a specialization of the above formalism in which each site is its own LFR, assuming that the movement of side chain i does not change the intrinsic properties of any other side chain, j, or the interaction between other pairs, jk. Eq. (2) then simplifies to:

$$\mu^\circ_M(k, \mathbf{c}, \mathbf{x}) = E_g(k) + \sum_i^{N_s} E_i(k, c_i) + P_i(k, c_i, x_i) + \frac{1}{2}\sum_{ij}^{N_s}{}' V_{ij}(k, c_i, c_j) + W_{ij}(k, c_i, c_j, x_i, x_j),$$

and eq. (3) becomes:

$$f(k, \mathbf{c}, \mathbf{x}) = Z^{-1}\exp\left\{-\beta\left[E_g(k) + \sum_i^{N_s}\left[E_i(k, c_i) + P_i(k, c_a, x_i) - \mu^\circ_{\text{H}^+}\nu_i(x_i) + 2.3\beta^{-1}\text{pH}\,\nu_i(x_i) + \frac{1}{2}\sum_j^{N_s}{}'\left(V_{ij}(k, c_i, c_j) + W_{ij}(k, c_i, c_j, x_i, x_j)\right)\right]\right]\right\} \quad (17)$$

where we have used the relation, $\mu_{\text{H}^+} = \mu^\circ_{\text{H}^+} - 2.3\beta^{-1}\text{pH}$. We can now introduce a conformation-dependent version of the usual notion of intrinsic $\text{p}K_a$. Tanford and Kirkwood[3] defined the intrinsic $\text{p}K_a$ of a site i, $\text{p}K_{\text{int},i}$, as the $\text{p}K_a$ that the site would have in a hypothetical state of the macromolecule in which all other sites are held in their neutral forms. In their formulation, this meant that the site–site interactions, W_{ij}, were zero in this hypothetical state. The same is true in the present case apart from subtleties regarding the dipole interaction of the neutral forms.* With the W_{ij} in eq. (17) set to zero, we can easily calculate the population ration of molecules with i protonated to those with i unprotonated:

$$\frac{[i\ \text{prot.}]}{[i\ \text{deprot.}]} = \frac{e^{-\beta P_i(k, c_i, \text{prot.}) + \beta\mu^\circ_{\text{H}^+} - 2.3\text{pH}}}{e^{-\beta P_i(k, c_i, \text{deprot.})}} = 10^{[-P_i(k, c_i, \text{prot.}) + P_i(k, c_i, \text{deprot.}) + \mu^\circ_{\text{H}^+}]/2.3RT - \text{pH}}$$

where ''prot.'' and ''deprot.'' are the specific values that the variable x_i can take on this case. In terms of intrinsic $\text{p}K_a$, this ratio is simply $10^{\text{p}K_{\text{int}} - \text{pH}}$. We therefore make the identification:

$$-2.3\beta^{-1}\text{p}K_{\text{int},i}(k, c_i) = P_i(k, c_i, i\ \text{prot.}) - P_i(k, c_i, i\ \text{deprot.}) - \mu^\circ_{\text{H}^+}$$

The right-hand side is just the standard chemical potential of protonated site i, in the conformer (k, c_i), in the absence of interactions with other sites. Exploiting the fact that $\nu_i(x_i)$ takes on the values 0 or 1 for the deprotonated or protonated

* The neutral state of a titrating residue may still have dipole or lower electrostatic moments, in which case, the W_{ij} terms can be set to zero only by considering the interaction with the neutral forms of other sites to be a contribution to site i's $\text{p}K_{\text{int}}$ (e.g., as part of the ''background'' terms of ref. 20). However, in the conformationally flexible case, this may interfere with the assumption that the motion of one site does not influence the $\text{p}K_{\text{int}}$ of another. In this case, one could, without loss of generality, define the $\text{p}K_{\text{int}}$ of a site as the $\text{p}K_a$ with all other sites having *all partial charges set to zero*, and include dipole interactions in (conformationally dependent) W_{ij}.

states, respectively, the exact expression [eq. (17)] can be written in terms of the pK_{int} rather than the P_i:

$$f(k,\mathbf{c},\mathbf{x}) = Z^{-1}\exp\left\{-\beta\left[E_g(k) + \sum_i^{N_s}\left[E_i(k,c_i) + 2.3\beta^{-1}(\mathrm{pH} - pK_{\mathrm{int},i}(k,c_a))\nu_i(x_i) + \frac{1}{2}\sum_j{}'\left(V_{ij}(k,c_i,c_j) + W_{ij}(k,c_i,c_j,x_i,x_j)\right)\right]\right]\right\} \quad (18)$$

Note that if conformational flexibility is omitted from the above expression, the standard formula for rigid, multisite H^+ ion binding is recovered.[5, 21]

Introducing the mean-field approximation with the cluster Γ then gives:

$$f_\Gamma(\mathbf{c},\mathbf{x}\mid k) = e^{\beta A_\Gamma(k)}\exp\left\{-\beta\left[\sum_{i\in\Gamma}E_i(k,c_i) + 2.3\beta^{-1}(\mathrm{pH} - pK_{\mathrm{int},i}(k,c_i))\nu_i(x_i) + \frac{1}{2}\sum_{i,j\in\Gamma}{}' V_{ij}(k,c_i,c_j) + W_{ij}(k,c_i,c_jx_i,x_j) + \sum_{\substack{i\in\Gamma\\ j\notin\Gamma}}\left\langle V_{ij}(k,c_i)\right\rangle_j + \left\langle W_{ij}(k,c_i,x_i)\right\rangle_j\right]\right\} \quad (19)$$

The computational complexity of IMC calculations for this case will depend on the total number of sites, their conformational variability, and the number of sites that occur in the clusters (which depends on the interactions and the cutoff energy used in forming the clusters). Suppose each of the N_s sites has c conformers, and each cluster, Γ, contains g sites. A single iteration of the IMC method requires N_s cluster calculations (one for each site). A cluster calculation involves a summation over all possible states of the g sites in the cluster, and because each site can have two protonation states and c conformers there are $(2c)^g$ states. For each state, the chemical potential expression must be evaluated and the exponential function must be calculated. The former is dominated by site–site interaction terms of which there are $(gc)^2$ within the cluster and $gc(N_s - g)$ between the cluster and the mean field. The computational complexity of a single iteration is then:

$$N_s(2c)^g\left[t_c(gc)^2 + t_{\mathrm{mf}}gcN_s + t_e\right]$$

where t_c is a characteristic time for processing interaction within the clsuter, t_{mf} is a time for cluster–mean-field interactions, and t_e is a time for evaluating the exponential function and accumulating the sum. The formal dependence on N_s of the above expression is linear to quadratic depending on whether the t_c or the t_{mf} term is more dominant. In more realistic situations, not all clusters have the same size, g, but the average size is expected to be independent of N_s if the density of sites on a molecule remains constant. However, the larger clusters may dominate the calculation, and the larger the molecule, the greater the chance that some unusually large clusters will appear. Therefore, the effective g may have some dependence on N_s, depending on the statistical properties of the site–site interactions with increasing molecular size. Finally, it is possible that the number of iterations required for convergence will depend on N_s. The actual scaling with N_s can only be determined by numerical tests, and will depend on the details of the implementation and the systems chosen.

One-Site Cluster Limit

The need for mobile clustering can be eliminated and a unique free energy functional can be retained by making the much more drastic approximation that all correlations between sites can be neglected so that the distribution function has the form:

$$f(k,\mathbf{c},\mathbf{x}) = f_g(k)\prod_i^{N_s}f_i(c_i,x_i\mid k)$$

(In the conformationally rigid case, this leads to the Tanford–Roxby approximation.[4, 5]) Putting this form of the distribution into a free energy functional expression and solving for the minima with respect to the f_i yields results analogous to eq.

(19), with Γ composed of only one site:

$$f_i(c_i, x_i \mid k) = e^{\beta A_i(k)} \exp\left\{ -\beta \left[E_i(k, c_i) + 2.3\beta^{-1}(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},i}(k, c_i))\nu_i(x_i) + \sum_{j \neq i} \left\langle V_{ij}(k, c_i) \right\rangle_j + \left\langle W_{ij}(k, c_i, x_i) \right\rangle_j \right] \right\}$$

The Lagrange multiplier expressions have the form of free energies for the canonical ensemble of a single site in the mean field of all the other sites:

$$e^{-\beta A_i(k)} = \sum_{c_i x_i} \exp\left\{ -\beta \left[E_i(k, c_i) + 2.3\beta^{-1}(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},i}(k, c_i))\nu_i(x_i) + \sum_{j \neq i} \left(\left\langle V_{ij}(k, c_i) \right\rangle_j + \left\langle W_{ij}(k, c_i, x_i) \right\rangle_j \right) \right] \right\}$$

Minimization of the free energy with respect to the global part of the distribution leads to an expression for the global distribution analogous to eq. (16):

$$f_g(k) = e^{\beta A_g} \exp\left\{ -\beta \left[E_g(k) + \sum_i^{N_s} A_i(k) - \frac{1}{2} \sum_{ij}^{N_s} \left\langle V_{ij}(k) \right\rangle_{ij} + \left\langle W_{ij}(k) \right\rangle_{ij} \right] \right\} \quad (20)$$

As in eq. (16) the mean interaction terms enter with a negative sign to cancel the double counting of interactions within the A_i. In contrast to eq. (16), this expression can actually be evaluated in practical calculations.

GLOBAL CONFORMATIONAL PROBLEM

As noted earlier, the general cluster mean-field expressions do not yield a tractable expression for f_g, the global part of the distribution function, and the mobile clustering approach does not allow for a unique free energy functional. The root of the latter problem is that the entropy term of a free energy functional involves expressions like $f \ln f$, that can only be decomposed into terms involving small groups of sites by neglecting specific sets of correlations, whereas the mobility of our clustering approach means that the correlations neglected are different in different phases of the calculation.

An alternative approach that does not completely solve the problem but may be useful in many practical situations involves integration over the binding isotherms.[22,23] Starting from eqs. (3)–(5), and expressing the dependence of the distributions on the ligand chemical potentials explicitly, the distribution function is:

$$f(k, \mathbf{c}, \mathbf{x}) = \frac{e^{-\beta(E_g(k) + A_g(k, \mu_1, \mu_2, \ldots))}}{\sum_k e^{-\beta(E_g(k) + A_g(k, \mu_1, \mu_2, \ldots))}}$$

The A_g values in this expression are the free energies for the semigrand canonical ensembles over the binding states and LFRs within the global conformations labeled by k, and are the logarithms of the terms in the summation over k in eq. (3). Their dependence on the ligand chemical potentials is contained in terms of the form:

$$\sum_i \cdots - \mu_1 \nu_{1i}(x_i) - \mu_2 \nu_{2i}(x_i) \cdots$$

in the exponentials within the sums over protonation states and local conformers. Differentiating with respect to the chemical potentials of the ligands gives:

$$\left(\frac{\partial A_g(k)}{\partial \mu_1} \right)_{\mu_2 \ldots} = \sum_i^{N_s} \langle \nu_{1i} \rangle_k = \langle \nu_1 \rangle_k$$

where $\langle \nu_{1i} \rangle_k$ is the mean number of ligands of type 1 bound to site i when the global conformer is held to k and all local flexibility and binding states are averaged over. These averages can be obtained from the IMC procedure. The A_g at one set of values of the ligand chemical potentials can then be related to the A_g at any other set of chemical potentials through integrals over the binding isotherms. In the case of only one type of ligand, the expression is:

$$A_g(k, \mu_1') - A_g(k, \mu_1'') = \int_{\mu_1''}^{\mu_1'} \langle \nu_1 \rangle_k \, d\mu_1$$

or, if the ligands in question are protons:

$$A_g(k, \mathrm{pH}') - A_g(k, \mathrm{pH}'') = -2.3\beta^{-1} \int_{\mathrm{pH}''}^{\mathrm{pH}'} \langle \nu_1 \rangle_k \, d\mathrm{pH}$$

Such expressions are useful if there is some value of the ligand potentials for which the A_g or their differences can be calculated explicitly. For example, at an extremely high pH, where all sites become deprotonated, the combinatorial problem may become simple enough to allow for a direct

calculation of the A_g. The integral expressions are also useful if there is some value of the chemical potential at which the ratio of the populations of the global conformers are known. For example, at the midpoint of a pH-dependent change between global conformers k and m one must have $E_g(k) + A_g(k, \mathrm{pH}_{\mathrm{mid}}) = E_g(m) + A_g(m, \mathrm{pH}_{\mathrm{mid}})$, and the integral formulas can be used to find differences in A_g at other pH values.[23]

Implementation and Test Methods

The IMC method just outlined clearly prevents the exponential growth of computational cost with the total number of sites seen in the exact formulation of the problem, but its cost still grows exponentially with the number of sites or LFRs in the clusters. The cluster size depends on the interaction energy threshold used to define the clusters. Raising the threshold reduces the cost, but increases the chance of significant errors in the results. Can a threshold be found that yields accurate results at acceptable cost when the method is applied to systems with energetics characteristic of interesting macromolecules? The test models and the implementation of IMC presented here are designed to answer this question.

DESCRIPTION OF TEST MODEL

The test problems are simple pH-titration problems with a one-to-one correspondence between proton binding sites and LFRs. The formulas for these problems have been presented in the subsection on pH titration with flexible sidechains. Although such systems have a simpler formalism than the general case, they are no less demanding in combinatorial terms. The number of binding states and the number of combinations of local conformers both grow exponentially with the total number of sites, and the cost of the IMC method is expected be exponential in the number of sites per clsuter. No global conformational change is included, because the potential methods of handling this problem are quite different and are not the main focus of the present study. The energetic characteristics of the test systems are chosen to mimic those likely to be found in proteins, and the local flexibility is chosen to mimic side-chain rotation. The most important parameters with regard to the complexity of the titration problems are the site–site interactions, which are regarded as distance-dependent electrostatic interactions between proton charges at the various sites. Finally, we use simple model geometries and interaction functions so that large numbers of test systems of various sizes can be generated randomly and calculations can be carried out on them rapidly.

The test macromolecule is a square surface on which site centers are placed at random, with the stipulation that no two site centers are closer than 3 Å. Half of the sites are chosen to be anionic with $\mathrm{p}K_{\mathrm{int}} = 4$ (characteristic of aspartic acid residues) and half are chosen to be cationic with $\mathrm{p}K_{\mathrm{int}} = 10$ (characteristic of lysine). The density of sites is controlled through the number of sites and the size of the square and, in most experiments, is set to one site per 90 Å^2, a value characteristic of real proteins, as shown in what follows. A surface-based model was chosen because most ionizable groups in proteins occur on the surface. For the present purpose, the most important characteristic of the tests systems (which would not be altered much by a small fraction of nonsurface sites) is the statistical distribution of site–site interactions that will influence the cluster sizes of the IMC method and the convergence of the Monte Carlo method. The parameters of the model were chosen to yield statistics similar to those found in a survey of real proteins (see ''Clustering in Real Proteins,'' subsection).

The local conformers of each site are represented by points chosen randomly within a 3-Å radius of the site center. To avoid unreasonably strong interactions, if the conformer point of one group is closer than 3 Å from the conformer point of another group, the interaction energy between them is taken to be that corresponding to 3 Å. The $\mathrm{p}K_{\mathrm{int}}$ values of the sites do not vary with conformation. A few tests in which the $\mathrm{p}K_{\mathrm{int}}$ values did vary were tried, but neither the cost nor the accuracy of the results were significantly affected (results not shown).

The site–site interaction functions $W_{ij}(c_i, c_j, x_i, x_j)$ are zero, positive, or negative, according to the protonation state and the anionic or cationic character of the interacting sites. The magnitude of the nonzero interactions depends on the distance $r(c_i, c_j)$ in angstroms, between the conformer point of site i in conformer c_i and the conformer point of site j in conformer c_j according to the inverse square law:

$$|W_{ij}| = \frac{332}{br^2(c_i, c_j)} \tag{21}$$

where the energy units are in kilocalories per mole. In all of the tests, $b = 4$. This form corresponds to an interaction with an ''effective dielectric constant'' of br at a temperature of 300 K. Such forms are often used in molecular mechanics calculations as a simplified way to represent the increase of solvent screening effects with distance.[24] An inverse-square law has been found to give a good account of the evolutionary optimization of electrostatic interactions in proteins.[25] The choice, $b = 4$, gives a reasonable fit to finite-difference solutions to the Poisson equation for calbindin,[1] and to an empirical function proposed for electrostatic interactions in proteins[26] in the 5- to 10-Å range, which is where the crucial cutoffs occur. On the other hand, comparisons to finite-difference calculations for lysozyme[21] suggest that $b = 4$ may overestimate the interactions in proteins, where most sites are well exposed to solvent (see also ref. 27).

The criterion for forming clusters in the IMC method is that a site j is included in the cluster centered on site i if any combination of conformers of i and j has an interaction energy that exceeds a specified cutoff ξ. The self-consistency test for stopping the outer iterative loop of the IMC method is that the mean of the absolute values of the differences in the individual site protonations between two consecutive iterations must be less than 0.00125.

SOFTWARE, DATA, AND COMPUTER

The generation of the test models through random site and conformer placement, and the solution of these test problems by the exact formulas [eqs. (3) and(4)], the IMC method, or a Monte Carlo method, have been integrated in a single computer program written in C++. Interactions are either calculated as needed from conformer point coordinates according to the inverse-square formula given earlier, or obtained from a tabulation of interactions prepared in advance; the former method is slightly more advantageous for larger systems, and the latter is more advantageous for smaller systems. The Monte Carlo method implemented is that described by Beroza and Case.[10] We have verified our implementation of the Monte Carlo method by comparison with the XMCTI program of Beroza. On small systems, our implementation achieves somewhat higher accuracy for the same number of steps compared with XMCTI. The CPU time used by our Monte Carlo implementation for larger systems is roughly one third that used by XMCTI and the scaling behavior is the same. Our computer program is available as freely modifiable and redistributable source code through the Bashford group World Wide Web page, http://www.scripps.edu/bashford.

Calculations were carried on a Silicon Graphics O2 workstation equipped with 64 megabytes of memory and a MIPS R10000 processor running at 175 MHz. The operating system was IRIX 6.3, and version 2.8 of the GNU C/C++ compiler was used. Execution timings were obtained using the TIMES system call.

A survey of 178 protein structures from the Protein Data Bank[28,29] was carried out to determine the distribution of cluster sizes produced using distance between ionizable groups as a cutoff criteria. The distances were chosen to correspond to the energy cutoffs used in the IMC tests through the inverse-square law. All aspartic acid, glutamic acid, tyrosine, lysine, arginine and histidine residues were considered ionizable. The group of 178 proteins is a subset of a ''representative selection'' of 200 proteins assembled by Boberg et al.,[30] from which we have excluded membrane proteins and proteins with incomplete atom listings for the titrating side chains. The survey was implemented by a simple program written in the NAB language, which is a C-like language specialized for macromolecular manipulations.[31]

Results

ACCURACY TESTS ON EXACTLY SOLUBLE SYSTEMS

The accuracy of the IMC method was assessed by carrying out calculations on systems small enough to be solved exactly. These systems had eight sites, each of which had two conformers; the density of sites was one per 90 Å^2, and the inverse-square-law parameter b was set to 4. For comparison, Monte Carlo calculations were also carried out for these systems. The Monte Carlo calculations used 10,000 steps, where a ''step'' is defined as a sweep through all of the sites; this is consistent with the original presentations of the Monte Carlo methods for titration calculations.[6]

Figure 1 shows how the mean absolute error in the degree of protonation varies with the cutoff energy ξ used in the IMC scheme. At the $\xi = 0$ limit, the approximation scheme becomes identical to the exact solution (because all sites are in the cluster). At large ξ values, where the IMC method

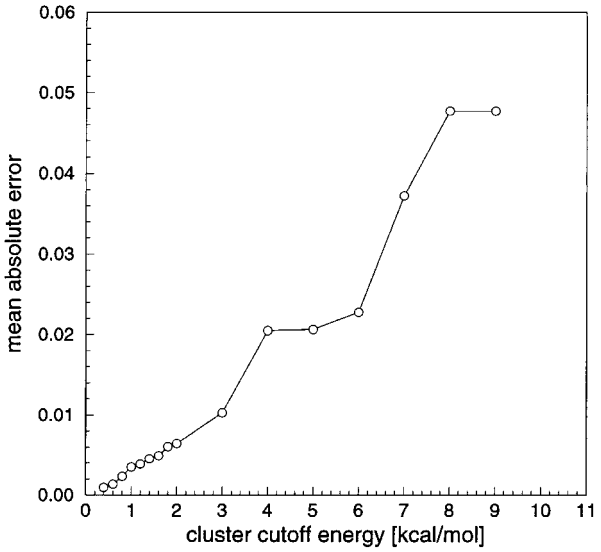

FIGURE 1. Mean absolute error in the IMC calculation of the protonation fraction of sites versus the interaction energy cutoff used as the clustering criteria of the IMC method. Twenty instances of the small system (see text) were randomly generated and solved at pH values ranging from -5 to 15 in intervals of 0.5. Error averaging ran over all eight sites of all 20 systems and, for a particular site and system, included only pH values at which the exact solution gave a protonation fraction of between 0.1 and 0.9. This excluded the tails of the titration curves, which would have artificially reduced the mean error.

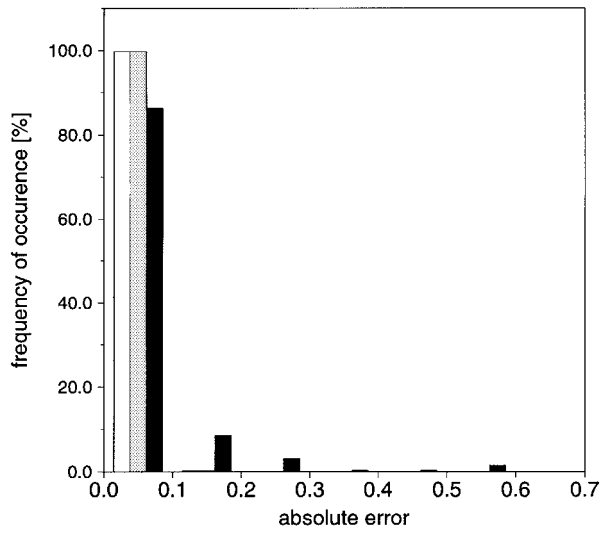

FIGURE 2. Ranges of error in a site's protonation fraction versus frequency of errors in those ranges for various methods: IMC with $\xi = 1$ kcal / mol (white bars); IMC with $\xi = 2$ kcal / mol (gray bars); and Monte Carlo with 10,000 steps (black bars). The test systems were as described in Figure 1. Errors are relative to the exact solution, and statistics were gathered over all sites of the 20 generated systems over a range of pH values, as for Figure 1.

goes over to the one-site cluster limit, the error rose significantly. For some test cases, calculations at or near the one-site cluster limit could not be completed because the iterative scheme did not converge to a self-consistent solution. The convergence problems of the closely related Tanford–Roxby approximation in situations of strong coupling have been noted previously.[5] When a cutoff of $\xi = 1$ kcal/mol was used, a mean absolute error of only 0.003 in protonation was obtained. A cutoff of $\xi = 2$ kcal/mol leads to a somewhat larger mean absolute error of 0.006. These two cutoff levels, both of which lead to quite reasonable accuracy in terms of average error, were selected for further study. The mean absolute error of the Monte Carlo method was 0.055, an order of magnitude larger than the error of the IMC scheme, but still small in absolute terms. Lowering number of steps from 10,000 to 1000 did not significantly increase the error of the Monte Carlo method (results not shown).

In multisite titration problems there can be pathological cases in which strong couplings or unexpected correlations cause approximation methods to fail, resulting in significant errors for a few cases even when the overall mean error is low. This issue is explored in Figure 2, which displays the frequency of errors of various sizes for several different methods. The IMC method with either $\xi = 1$ or 2 kcal/mol proved very reliable, with 99.75% of the calculated protonation fractions having errors of less than 0.1, and no errors greater than 0.2. The Monte Carlo method is considerably less reliable. The results of a 10,000-step calculation had errors of less than 0.1 only 86% of the time, whereas errors of greater than 0.2 occurred 5% of the time. As with the mean average error, lowering the number of Monte Carlo steps to 1000 did not significantly change the distribution of errors.

COMPUTATIONAL COST AND LARGER SYSTEMS

Calculations have been carried out on randomly generated model systems with two conformers per site and the number of sites ranging from 8 to 150. The density of sites was one per 90 Å^2 in all models. Figure 3 shows that the average number of sites in a cluster for the IMC method was sensitive to the cutoff energy ξ, but nearly inde-

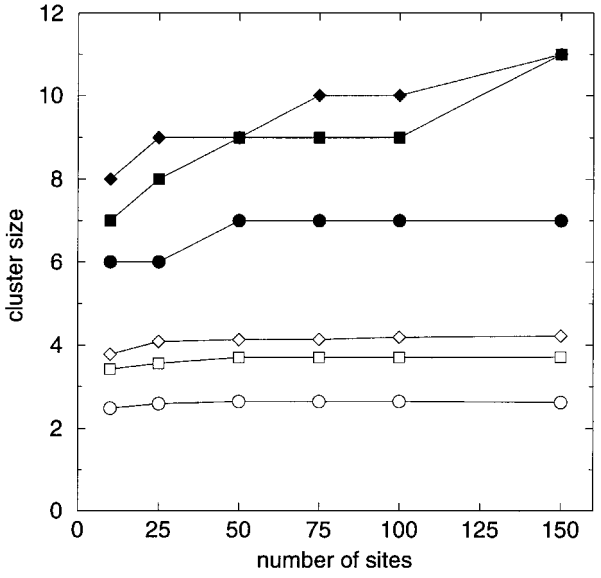

FIGURE 3. The mean cluster size in the IMC method (open symbols) and the maximum cluster size occurring in each test run (closed symbols) versus the total number of sites, N_s. Results included for tests using cutoff energies: $\xi = 1$ kcal / mol (diamonds); $\xi = 1.2$ kcal / mol (square); and $\xi = 2$ kcal / mol (circles).

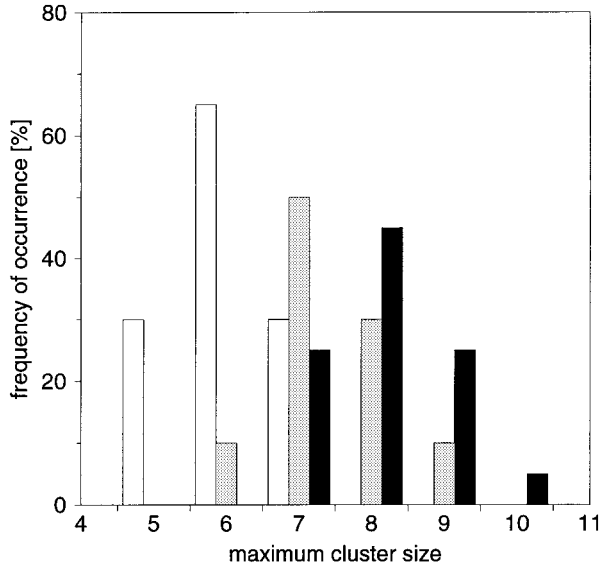

FIGURE 4. The frequency of occurrence of maximum cluster sizes within a particular test system. Statistics over a set of 20 randomly generated test systems with 100 sites each. Black bars: $\xi = 1$ kcal / mol; gray bars: $\xi = 1.2$ kcal / mol; and white bars: $\xi = 2$ kcal / mol.

pendent of the total number of sites. However, the complexity analysis of the IMC method suggests that computation time may be dominated by the largest clusters occurring in the system; therefore, the figure also shows the average over test systems of the maximum cluster size occurring in that test system, as a function of the number of sites. This quantity had a weak dependence on the number of sites, which might have been expected because a larger number of sites increases the chance of placing a larger group of them close together, even as the average density is held constant. The histograms of Figure 4 provide a more detailed view of the maximum cluster sizes occurring in the test systems with $N_s = 100$. At $\xi = 2.0$ kcal/mol, the most common maximum cluster size is 6, whereas at $\xi = 1.0$, the most common maximum cluster is 8, which is much more computationally demanding, because each additional site in a cluster increases computation time by roughly a factor of four. Even a slight increase in ξ, to 1.2 kcal/mol, which approximately shifts the cluster size histogram down by one site, can potentially give significant savings in computation.

The average number of iterations required by the IMC method with $\xi = 1$ kcal/mol is less than six for systems with up to 100 sites, and is only weakly dependent on the number of sites N_s (Fig. 5). The maximum number of iterations needed by any of the 20 test cases at each N_s is more erratic, but never exceeds 14. As ξ is increased, the number of iterations needed for convergence also increases, but as the CPU-time results discussed in

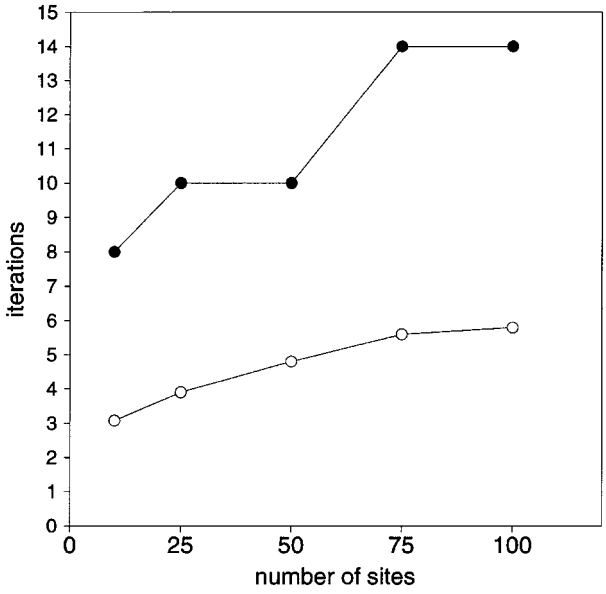

FIGURE 5. Mean (open circles) and maximum (closed circles) number of iterations required for convergence of the IMC method with $\xi = 1$ kcal / mol, versus number of sites, N_s, in the test systems. Averages taken over 20 test systems at each N_s value.

what follows demonstrate, the time savings due to smaller clusters far outweigh the cost of more iterations, provided ξ is not set so high as to cause failure of convergence.

The CPU timings shown in Figure 6 demonstrate that the IMC method was feasible for large problems with the 100-site problem requiring approximately 10 minutes of CPU time when the cutoff of $\xi = 1$ kcal/mol is used. Between $N_s = 8$ and 75, the computational cost increased, as expected, but from 75 to 100 sites there was a slight drop in cost, which was an artifact of statistics dominated by a few test systems that contained a few large clusters. The times required for individual test cases for any particular N_s varied dramatically, and was very sensitive to the maximum cluster size for that test case. For example, at $N_s = 100$, the cases with a maximum cluster size of 7 averaged 124 seconds run, whereas those with maximum cluster sizes of 10 averaged 2500 seconds run.

Raising the cutoff energy to $\xi = 2.0$ kcal/mol dramatically increased the speed, but for some tests the iterative method did not converge at this cutoff: 15% of the tests at $N_s = 50$ and $N_s = 75$, and 30% of the tests at $N_s = 100$, failed to converge. A more cautious raising of the cutoff energy, from $\xi = 1.0$ to 1.2 kcal/mol, results in quite substantial savings in CPU time, and does not result in any nonconvergent runs. The Monte Carlo method's cost grows quadratically with the number of sites, in accordance with previous results,[6] and the overall CPU time requirements are much greater than those for the IMC method even with the more stringent and costly cutoff value. However, the Monte Carlo method's cost is proportional to the number of steps (here 10,000), and the accuracy tests (see later) suggested that a significantly smaller numbers of steps do not substantially worsen the accuracy of this method.

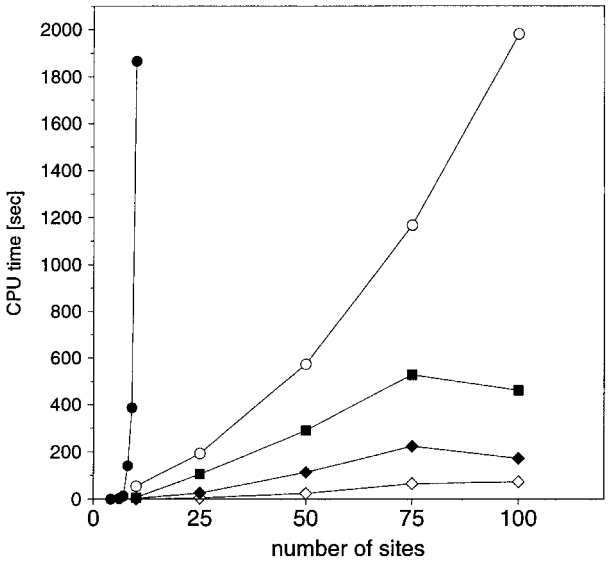

FIGURE 6. Computer CPU time versus total number of sites, N_s, for the following different methods: exact (closed circles); 10,000-step Monte Carlo (open circles); IMC $\xi = 1$ kcal/mol (closed squares); IMC $\xi = 1.2$ kcal/mol (closed diamonds); and IMC $\xi = 2$ kcal/mol (open diamonds). Averages are taken over 20 test systems at each N_s point for each method. Some of the IMC $\xi = 2$ kcal/mol runs did not converge and are not included in the averaging (see text).

It is not possible to determine the absolute accuracy of either the IMC or Monte Carlo methods for large systems, because the exact solutions are not calculable in practice, but the small-system result that the IMC method with $\xi = 1.0$ kcal/mol was quite accurate and reliable suggests that IMC with this ξ value is a suitable basis for estimating the accuracy of other methods. The difference between the IMC-with-$\xi = 1$ kcal/mol calculations of protonation fractions and those of IMC with $\xi = 1.2$ kcal/mol, the convergent subset of the $\xi = 2.0$ kcal/mol runs, and the nonconvergent subset of the $\xi = 2.0$ kcal/mol runs for $N_c = 50$ are shown in Figure 7. The $\xi = 1.2$ kcal/mol results showed almost no difference from the $\xi = 1.0$ kcal/mol results, so there was no sacrifice in accuracy for this raising of ξ, even though it gave a substantial reduction in computational cost. However, at $\xi =$ 2.0, there were substantial differences for both the convergent and nonconvergent cases. If the difference from the $\xi = 1.0$ kcal/mol calculation is taken as an indication of error, this means that lack of convergence can be taken as an indication that ξ has been set too large and errors are likely, but convergence alone cannot be taken as an indication that the results are accurate. The differences between Monte Carlo and $\xi = 1.0$ kcal/mol results were quite similar to the differences between Monte Carlo and the exact solution for small systems (compare Fig. 2), suggesting that the accuracy of the Monte Carlo method is substantially worse than even the nonconvergent IMC runs. Monte Carlo calculations with only 4000 steps, rather than 10,000, gave almost identical histograms (not shown).

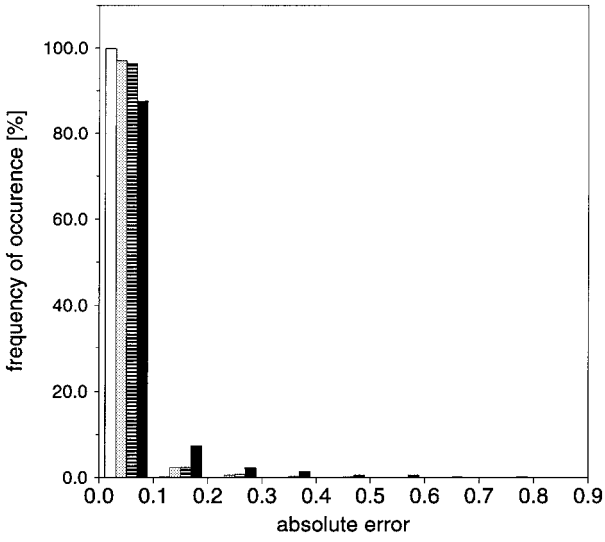

FIGURE 7. Distribution of absolute differences in a site's protonation fraction between those calculated by IMC with $\xi = 1$ kcal / mol, and those calculated by other methods: IMC with $\xi = 1.2$ kcal / mol (white bars); convergent subset of IMC with $\xi = 2.0$ kcal / mol (light gray bars); nonconvergent subset of IMC with $\xi = 2.0$ kcal / mol (hatched bars), and Monte Carlo with 10,000 steps (black bars). Statistics are taken over 20 test systems with $N_s = 50$ with difference sampling and averaging by the methods described for Figures 1 and 2.

CLUSTERING IN REAL PROTEINS

In the survey of structures from the Protein Data Bank, two clustering criteria were used: a cutoff distance $r_c = 9$ Å, which corresponds to $\xi \simeq 1$ kcal/mol with the inverse square law, eq. (21), and $b = 4$; and $r_c = 6.5$ Å, which corresponds to $\xi \simeq 2$ kcal/mol. The distribution of the average and maximum number of sites per cluster is shown in Figure 8. Across the entire set of proteins, the average cluster size at $r_c = 9$ Å, is 3.81, whereas the average size of the largest cluster in a protein (which is an indicator of the upper limit of the cost of IMC calculations for that protein) is 7.2. With the cutoff $r_c = 6.5$ Å, the average cluster size is 2.37, and the average size of the largest cluster in the protein is 4.71. Only 25% of the proteins have any cluster of more than eight sites at $r_c = 9$ Å, and at $r_c = 6.5$ Å no clusters of more than eight sites occurs. If there are two conformers per site, the number of states involved in a cluster of eight sites is $4^8 = 65{,}536$, which is easily manageable.

Randomly generated test systems with a density of one site per 90 Å^2 have clustering properties similar those found in the set of real proteins, as

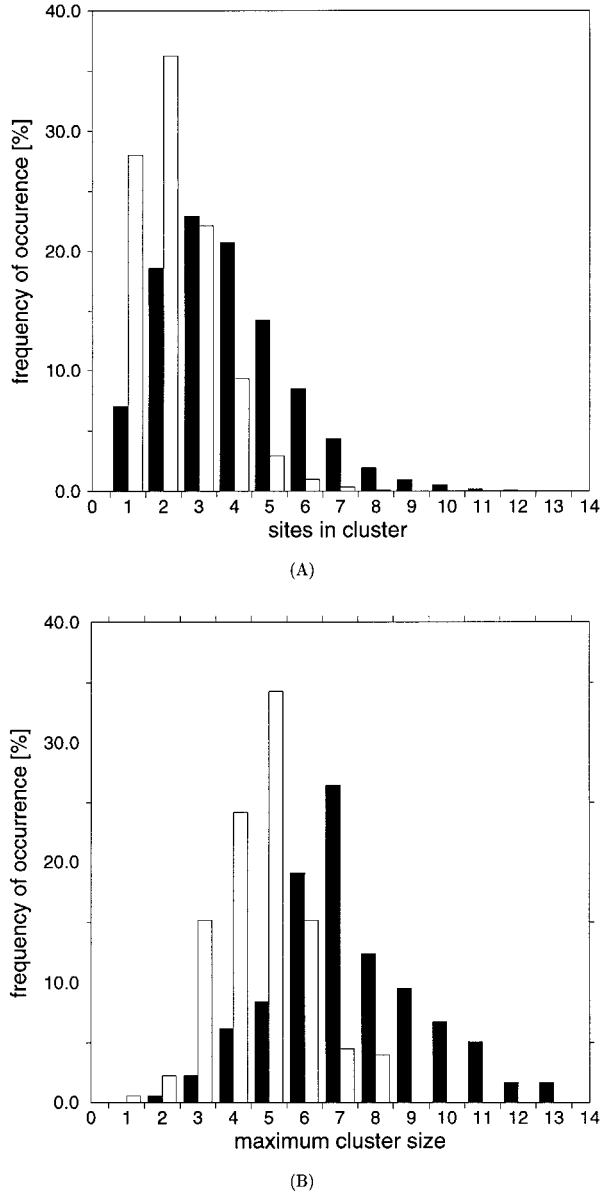

FIGURE 8. Clustering in real proteins using clustering criteria of 9 Å (black bars), or 6.5 Å (white bars). (A) Frequency of occurrence of clusters of given size across the full set of protein structures. (B) Frequency of occurrence of proteins for which the largest cluster is of a given size.

can be seen by comparing the average and average maximum cluster sizes quoted earlier with Figure 3, or Figure 4 with Figure 8B.

Discussion

We have presented a formalism that includes the genuine complexity of the binding of ligands

to multiple sites in conformationally flexible molecules, while allowing considerable scope for approximation schemes to make large problems tractable. The locally flexible region (LFRs) provide a way of expressing conformational states with a relatively small number of indices (the c_a), that each have a small number of possible values, but whose combinatorics captures some of the conformational complexity of macromolecules with many degrees of freedom. The global conformational index k provides a means of including conformational variation that cannot be expressed in terms of the LFRs, so that, in principle, even the most conformationally complex situation can be described. Although the conformational complexity of a protein molecule would still lead to an intractable problem if expressed fully, it should be possible to formulate more limited, but still interesting, problems that are solvable. For example, myoglobin is known to undergo a transition from the folded state to a ''molten globule'' in which some helices are lost, under mildly acidic conditions.[32] The folded state could be represented by a single global conformer k, and a set of LFR variations representing side-chain conformers. The molten globule might be represented by an ensemble of several hundred global conformers supplemented by LFR variation again representing side-chain motions. Solving the combinatorial part of this problem could be done by several hundred IMC runs (one for each global conformers), a task that could be accomplished in hours or days of CPU time, and is trivially parallelizable.

In terms of the physical model, the key to almost any computationally tractable approach is the assumption of pairwise additive interactions in the functional form of the macromolecular chemical potential [e.g. eq. (2)]. This arises naturally from any linear response theory such as semimacroscopic electrostatics,[33] or Gaussian fluctuation models.[34] However, the applicability of linear response theory may be limited to cases in which the macromolecule undergoes only small fluctuations about an equilibrium position. The inclusion of multiple conformers through a formalism such as ours is then complementary, because each discrete ''conformational state,'' as expressed by the **c** and k variables, can be taken to represent a small region of conformational space within which the linear response assumption is valid. Nonlinearity in the interdependence of ligand-binding events then arises through the explicit conformational changes, even while the underlying theory of interactions is linear.

IMC METHOD

We have shown how approximation schemes can be constructed through neglect of correlations and, in particular, we have presented an iterative mobile clustering (IMC) method that should be generally applicable to the combinatorial problem that arises from multiple-site binding and multiple-region conformational variation. The crux of the difference between IMC and a mean-field method without mobile clustering, is that, although approximations are formulated in terms of multisite distribution functions, we are interested only in the one-site distribution functions, eqs. (13) and (14), because the quantities that are ultimately to be calculated are one-site averages or their sums —for example, the protonation fraction of a site or the total protonation of a molecule.

The tests on exactly solvable systems show that the IMC method gives very good accuracy when used with cutoff energies that, for larger systems, lead to clusters small enough for rapid computation. The method is considerably more accurate than Monte Carlo for the exactly solvable systems, and does not suffer nearly so much from pathological cases that lead to instances of large errors at a few sites. Although comparisons to exact solutions are not possible for large systems, the test results for large systems are consistent with the hypothesis that the large-system accuracy of both the IMC and Monte Carlo methods are similar to their small-system accuracy (compare Figs. 2 and 7). For all system sizes, the IMC method is much faster than the Monte Carlo method with 10,000 steps, but it may be possible increase the speed of the latter simply by reducing the number of steps without much effect on accuracy.

Because the various Γ values in the IMC scheme are set up according to a cutoff criteria that relates only to the *central* site of the cluster, there is the possibility of error if significant correlations exist between the noncentral residues inside the cluster and those outside. It is expected that the effects of this on the accuracy of the overall scheme will be minimal because the cluster calculation is used only for determining the one-site distribution of the central site. The tests presented here bear out this expectation for protein-like systems, provided that the cutoff is sufficiently low. In principle, however, the fact that direct interactions fall below some cutoff is no guarantee of lack of correlation. One can imagine Ising-model-like systems in which chains of strong interactions between nearest neighbors lead to long-range correlation that would

cause the IMC method, or many other mean-field methods, to fail. Monte Carlo approaches to such problems would also encounter difficulties, such as becoming trapped in a high-energy state from which all local moves lead to still higher energies, whereas only a global move involving the simultaneous change of many sites' states would lead to lower energies. Although our tests suggest that the IMC scheme will be very useful for protein-like systems, the possibility for pathological cases still exists, and the methods should be used with caution, especially in systems with many strong site–site interactions.

The finding that accurate results can be obtained using the IMC method with acutoff of $\sim 2RT$ ($\xi = 1.2$ kcal/mol) may seem surprising. There are two points that explain this result. First, large coupling does not necessarily lead to correlation in protonation fraction if the sites do not titrate in the same pH range.[5] Second, the quantity being calculated, protonation fraction, does not depend directly on the two-site distribution functions, so any influence of correlations will be indirect. Calculations of quantities depending on pair distributions, such as the average value of a site–site interaction, would require different standards of accuracy, and probably a significant modification of any mobile clustering scheme.

LIMITS AS TO LOCAL FLEXIBILITY

It is necessary either that the number of conformers of an LFR is rather limited (as in the test cases here) or that the number of LFRs is small (significantly less than the number of titrating sites in a typical protein), otherwise the combinatorics would lead to intractable problems even within the clusters. Too many local conformers may also cause problems for the part of the problem not considered here—the problem of finding all the necessary one-site intrinsic terms and two-site interaction terms. If a macromolecule has N_s sites, each of which has n_c conformers (assuming a one-to-one correspondence between sites and LFRs), the number of intrinsic terms needed is proportional to $N_s n_c$, and the number of interaction terms is proportional to $(N_s n_c)^2/2$. An example of excessive local conformational flexibility is provided by our previous work[9] in which 11 residues of lysozyme were each allowed 36 possible conformers. In the present formulation, this would lead to a need to tabulate over 70,000 site–site interactions, and the combinatorics of the conformational states would be prohibitive for more than three sites in a cluster. In that previous work, we reduced the problem of calculating energy terms by performing explicit calculations of only the 11×36 intrinsic terms, and using an average of a small subset of all possible interactions for the site–site interactions. The conformational combinatorial problem was avoided altogether by applying conformational averaging only to the calculation of intrinsic pK values, rather than to protonation states. A valuable insight from that work was that only a small number of conformers made a significant contribution to the final results, suggesting that, with suitable prescreening, the numbers of conformers per site could be greatly reduced.

STATIC CLUSTERING METHODS

A possible alternative to the IMC method used here would be a static clustering scheme in which the molecule is divided into several clusters, each small enough for explicit enumeration of states within the cluster, and the cluster boundaries would not change in the course of the mean-field iterations. A method of this kind was described by Gilson[7] for the single conformer case—and, in the limit of one site per cluster, this becomes the Tanford–Roxby method.[4] The advantage of static clustering is that one has a functional form for the distribution function and the free energy functional that is consistent in the sense that it does not change according to which site one is ''looking at'' in some phase of the calculation. This allows one to use expressions similar to eqs. (16) or (20) in the global conformational problem, or any other application involving the ligand-activity dependence of the macromolecular free energy. The problem with static clustering is that it requires finding clusters of sites, such that the sites in one cluster have only weak interactions with sites in all other clusters. This implies either that the sites fall into relatively isolated groups geometrically, or that there are very few nonweak interactions, or that the cutoff criteria for a ''weak'' interaction is set rather high, sacrificing accuracy.

FUTURE DIRECTIONS

We have implemented IMC in a computer program specialized for the test model described here, and made it available as we have described. We hope to develop this program into an implementation of the more general formulation and, in the process, to write it in a more easily extendable,

object-oriented style. It will also be incorporated into this laboratory's existing MEAD software suite for macromolecular electrostatic calculations,[35,36] which is also freely available from the Bashford group Web page. We hope that interested members of the community will participate in this process by downloading, using, modifying, sharing, and commenting on this software.

As noted earlier, other methods may in some cases have significant advantages over IMC, particularly fixed clustering, but may not be as widely applicable. It would be desirable to develop hybrid methods that would use fixed clusters where possible, and mobile clustering elsewhere. Methods that more aggressively detect and exploit lack of correlation are also possible. For example, the present method clusters sites together if their interactions are strong; however, if they titrate at very different values of ligand activity, they will be uncorrelated in spite of strong coupling.[5]

We have alluded to the problem of prescreening local conformers as if it were prior to the combinatorial problem dealt with here and, logically, it is; however, it may be useful in practice to move flexibly between these problems. For example, a rough calculation of intrinsic terms and interactions using a rapid approximation may at first yield a number of local conformers that is somewhat too large. An IMC calculation with high cutoffs would be tractable and, although its accuracy would be limited, it could provide information on which conformers are never significantly populated and could therefore be removed from subsequent calculations.

References

1. Spassov, V.; Bashford, D. Prot Sci 1998, 7, 2012–2025.
2. Linderstrøm-Lang, K. Comptes Lab Carlsberg 1924, 15, 1–29.
3. Tanford, C.; Kirkwood, J. G. J Am Chem Soc 1957, 79, 5333–5339.
4. Tanford, C.; Roxby, R. Biochemistry 1972, 11, 2192–2198.
5. Bashford, D.; Karplus, M. J Phys Chem 1991, 95, 9556–9561.
6. Beroza, P.; Fredkin, D. R.; Okamura, M. Y.; Feher, G. Proc Natl Acad Sci USA 1991, 88, 5804–5808.
7. Gilson, M. K.Proteins 1993, 15, 266–282.
8. Yang, A.-S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. Proteins 1993, 15, 252–265.
9. You, T.; Bashford, D. Biophys J 1995, 69, 1721–1733.
10. Beroza, P.; Case, D. A. J Phys Chem 1996, 100, 21056–20163.
11. Alexov, E. G.; Gunner, M. R. Biophys J 1997, 74, 2075–2093.
12. Ripoll, D. R.; Vorobjev, Y. N.; Liwo, A.; Vila, J. A.; Scheraga, H. A. J Mol Biol 1996, 264, 770–783.
13. Zhou, H.-X.; Bijayakumar, M. J Mol Biol 1997, 267, 1002–1011.
14. Baptista, A. M.; Martel, P. J.; Petersen, S. B. Proteins 1997, 27, 523–544.
15. van Vlijmen, H. W.T.; Schaefer, M.; Karplus, M. Proteins 1998, 33, 145–158.
16. Winn, J. S. Physical Chemistry; New York: Harper Collins, 1995.
17. Chaikin, P. M.; Lubensky, T. C. Principles of Condensed Matter Physics; Cambridge University Press: Cambridge, UK, 1995.
18. Dimitrov, R. A.; Crichton, R. R. Proteins 1997, 27, 576–596.
19. de Groot, S. R.; Mazur, P. Non-Equilibrium Thermodynamics; North-Holland: Amsterdam, 1962.
20. Bashford, D.; Gerwert, K. J Mol Biol 1992, 224, 473–486.
21. Bashford, D.; Karplus, M. Biochemistry 1990, 29, 10219–10225.
22. Wyman, J. Adv Prot Chem 1964, 19, 223–286.
23. Yang, A.-S.; Honig, B. J Mol Biol 1994, 237, 602–614.
24. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187–217.
25. Spassov, V. Z.; Karshikoff, A. D.; Ladenstein, R. Prot Sci 1994, 3, 1556–1569.
26. Warshel, A.; Russel, S. T.; Churg, A. K. Proc Natl Acad Sci 1985, 81, 4785–4789.
27. Matthew, J. B.; Gurd, F. R. N. Meth Enzymol 1986, 130, 413–436.
28. Abola, E. E.; Sussman, J. L.; Prilousky, J.; Manning, N. O. In: Carter, C. W., Jr.; Sweet, R. M., eds. Methods in Enzymology, Vol. 277; Academic: San Diego, 1997; p 556.
29. Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. In: Allan, F. H.; Bergerhoff, G.; Sieves, R., eds. Crystallographic Databases—Information Content, Software Systems, Scientific Applications, Data Commission of the International Union of Crystallography: Bonn, 1987; p 107.
30. Boberg, J.; Salakoski, T.; Vihinen, M. Prot Eng 1995, 8, 501–503.
31. Macke, T. J.; Case, D. A. In: Leontis, N. B.; SantaLucia, J., Jr., eds. Molecular Modelling of Nucleic Acids (ACS Symposium Series No. 682); American Chemical Society: Washington, DC, 1998; p 379.
32. Eliezer, D.; Yao, J.; Dyson, H. J.; Wright, P. E. Nature Struct Biol 1998, 5, 148–155.
33. Honig, B.; Nicholls, A. Science 1995, 268, 1144–1149.
34. Del Buono, G. S.; Figueirido, F. E.; Levy, R. M. Proteins 1994, 20, 85–97.
35. Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E., Biochemistry 1993, 32, 8045–8056.
36. Bashford, D. In: Ishikawa, Y.; Oldehoeft, R. R.; Reynders, J. V. W.; Tholburn, M., eds. Scientific Computing in Object-Oriented Parallel Environments (Lecture Notes in Computer Science); ISCOPE97, Springer: Berlin, 1997; p 233.